(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 April 2002 (25.04.2002)

PCT

(10) International Publication Number
WO 02/033893 A3

(51) International Patent Classification[7]: H04L 12/26

(21) International Application Number: PCT/US01/32309

(22) International Filing Date: 17 October 2001 (17.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/241,450	17 October 2000 (17.10.2000)	US
60/275,206	12 March 2001 (12.03.2001)	US
09/903,423	10 July 2001 (10.07.2001)	US
09/903,441	10 July 2001 (10.07.2001)	US
09/923,924	6 August 2001 (06.08.2001)	US
09/960,623	20 September 2001 (20.09.2001)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	09/960,623 (CIP)
Filed on	20 September 2001 (20.09.2001)
US	09/923,924 (CIP)
Filed on	6 August 2001 (06.08.2001)
US	09/903,423 (CIP)
Filed on	10 July 2001 (10.07.2001)
US	09/903,441 (CIP)
Filed on	10 July 2001 (10.07.2001)
US	60/275,206 (CIP)
Filed on	12 March 2001 (12.03.2001)
US	60/241,450 (CIP)
Filed on	17 October 2001 (17.10.2001)

(71) Applicant *(for all designated States except US)*: **ROUTESCIENCE TECHNOLOGIES, INC.** [US/US]; 167 2nd Avenue, San Mateo, CA 94401 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FEICK, Wayne Allen** [CA/US]; 930 Grand Street, Redwood City, CA 94061 (US). **FINN, Sean,** [US/US]; 1533 Escondido Way, Belmont, CA 94002 (US). **KARAM, Mansour,** [LB/US]; 707 Continental Circle, #421, Montain View, CA 94040 (US). **LLOYD, Michael A.** [US/US]; 160 Arundel Road, San Carlos, CA 94070 (US). **MADAN, Herbert, S.** [US/US]; 347 Blackfield Drive, Tiburon, CA 94920 (US). **MCGUIRE, James, G.** [US/US]; 2312 Gough Street, San Francisco, CA 94019 (US). **VILLAVERDE, Jose-Miguel,** Pulido [ES/US]; 1020 Bryant Street, Palo Alto, CA 94301 (US). **BALDONADO, Omar,** [US/US]; 700 Alester Avenue, Palo Alto, CA 94303 (US).

(74) Agent: **SUZUE, Kente**; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report:
23 January 2003

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/033893 A3

(54) Title: METHOD AND APPARATUS FOR COMMUNICATING DATA WITHIN MEASUREMENT TRAFFIC

(57) Abstract: Methods and apparatuses for communicating data within measurement traffic are described. Embodiments that send, receive and both send and receive are described. Some embodiments are described that compute statistics based at least partly on measurement traffic. Some embodiments are described that communicate computed statistics within measurement traffic. Some embodiments are described that rank and select paths based at least in part on computed statistics.

10/070339

PCT/US01/32309

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION

(PCT Rule 61.2)

To:

Commissioner
US Department of Commerce
United States Patent and Trademark Office, PCT
2011 South Clark Place Room CP2/5C24
Arlington, VA 22202
United States of America

in its capacity as elected Office

Date of mailing (day/month/year) 13 February 2003 (13.02.03)	
International application No. PCT/US01/32309	**Applicant's or agent's file reference** 24717-715
International filing date (day/month/year) 17 October 2001 (17.10.01)	**Priority date** (day/month/year) 17 October 2000 (17.10.00)
Applicant FEICK, Wayne, Allen et al	

1. The designated Office is hereby notified of its election made:

[X] in the demand filed with the International Preliminary Examining Authority on:

14 May 2002 (14.05.02)

[ ] in a notice effecting later election filed with the International Bureau on:

2. The election [X] was

[ ] was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland**	Authorized officer A.NICKITAS-E.(Fax 338-8995)
Facsimile No. (41-22) 338.89.95	Telephone No. (41-22) 338 9443

3/prts

METHOD AND APPARATUS FOR COMMUNICATING DATA WITHIN MEASUREMENT TRAFFIC

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to the field of networking. In particular, the invention relates to communicating data within measurement traffic.

Description of the Related Art

Internetworks such as the Internet provide a best-effort service and do not reserve resources for a path. Hence, performance characteristics of the path such as delay, jitter and loss can change over time due to routing changes, congestion, and lack of connectivity, and therefore it is important to being able to measure them. There are several tools available to measure the performance characteristics of a path:

Ping uses ICMP packets to measure reachability and round trip delay from a source host to a remote host.

Traceroute detects common reachability problems such as routing loops and network black holes by sending ICMP packets from a source host to a destination host, and by receiving ICMP responses from intermediate routers along the path between the source host and the remote host. Each intermediate router in the path decrements the TTL value stored in the header of an ICMP packet by one; when the TTL field expires (reaches the value zero) in a router, the router does not forward the packet towards the destination host. Instead, it returns the ICMP to the source host responding with a Time Exceeded response. By starting with an initial TTL value of 1 and gradually incrementing the TTL field in successive ICMP packets, the source host is able to receive an ICMP response from all the routers in the path. Traceroute also computes the round trip time of each ICMP packet, hence being able to determine the round trip delay between the source host and intermediate routers.

Pathchar measures congestion of a path by estimating performance characteristics of each node along a path from a source to a destination. Pathchar also leverages the ICMP protocol's Time Exceeded response to packets whose TTL has expired. By sending a series of UDP packets of various sizes to each hop, pathchar uses knowledge about earlier nodes and the round trip time distribution to this node to assess incremental bandwidth, latency, loss, and queue characteristics across the link connected to this node.

These tools are mainly used for troubleshooting purposes. A more formal attempt to measure performance characteristics of Internet paths is being developed by the IP Performance Metrics (IPPM) working group of the Internet Engineering Task Force (IETF). The IPPM working group has specified a general framework for measuring performance characteristics of a path, including specifications for clock synchronization and for size, number and inter-transmission time of measurement packets. The IPPM working group has also specified specific performance metrics for one-way delay, one-way inter-packet delay variation, and one-way loss, among others. The goal of the IPPM measurement framework is to allow service providers and other network providers to develop and operate and inter-operable measurement infrastructure, for performance and billing purposes, among other purposes.

However, even if this measurement infrastructure is in place, a way to communicate measurements and performance characteristics of measured paths to appropriate points of the network where decisions based on those performance characteristics can be made, is needed. In addition, this communication should be efficient, i.e., it should minimize the amount of bandwidth consumed.

SUMMARY OF THE INVENTION

The invention includes methods and apparatuses for communicating data within measurement traffic. Some embodiments of the invention will consist of a sender of measurement packets. Some embodiments of the invention will consist of a receiver of measurement packets. Some embodiments of the invention will consist of both a sender and a receiver of measurement packets.

In some embodiments of the invention, measurement packets will traverse one or more paths traversing at least a portion of an internetwork.

In some embodiments of the invention, the measurement packet will include information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet. In some embodiments, the measurement packet sizes and times between measurement packets will simulate the traffic pattern of one or more applications such as -- by way of a non-limiting example -- voice and video.

In an embodiment that includes a receiver of measurement packets, measurement statistics may be computed that are at least partly recomputed with the arrival of each measurement packet. This computation may include at least one of a moving average, an average based on the Robbins-Moro estimator, a window-based average, and a bucket-based average.

In some embodiments of the invention, the measurement packets will contain data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

In an embodiment wherein the data includes measurement statistics, the measurement statistics may include one or more of delay, jitter and loss.

Some embodiments of the invention will contain clocks synchronized by GPS, IRIG, NTP or NIST. Some other embodiments will use unsynchronized clocks and will adjust for clock skew and drift by performing computations on the measurement data.

In some embodiments of the invention, paths will be selected based at least in part on at least one of: one or more of the measurement statistics from the measurement packet and one or more of the computed statistics.

These and other embodiments are described further herein.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 shows some possible embodiments of devices that are communicating with each other, for example sending and receiving measurement packets.

Fig. 2 shows one specific detailed embodiment of two devices, where each device is sending and receiving measurement packets as well as selecting a subset of paths.

Fig. 3 shows an embodiment with more than two devices that are sending and receiving measurement packets to obtain measurements of performance characteristics of paths and to communicate measurements statistics about those paths.

DETAILED DESCRIPTION

Measurement Packets

A measurement packet is a packet sent by a sender over an internetwork that includes information necessary for the receiver of the packet to compute measurements of the performance characteristics of the path the packet has traversed over that internetwork. The information includes information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet; and data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

In some embodiments of the invention, the information included in the measurement packet to compute measurements includes at least one of a timestamp of a sending time of the packet and a number to identify the packet by itself and/ to identify the relative position of the measurement packet in a sequence of measurement packets,

In some embodiments of the invention, the measurement packet is implemented using the following data structure:

```
struct MeasurementHeader {

  /**
   * A generation number.  This value represents when the
   * sender began sending.  This value is a standard Unix
   * timestamp that seconds since Jan 1, 1970 UTC.
   **/
  uint32_t mGeneration;

  /**
  * A sequence number for the packet. This increments each
  * time a packet is sent and rolls over when 16 bits is
   * exceeded.
   **/
```

```
    uint16_t mSequence;

    /**
     * The IP address the packet is sent to.
     **/
    uint32_t mDstAddr;

    /**
     * The send timestamp for this packet.
     **/
    uint64_t mSendTime;
};
```

The mGeneration field is used to detect when a sending process has started a new session. This field is used by the receiver to determine that a discontinuity in the stream's sequence numbers is the result of a sender restart, rather than due to large network latencies, duplicate packets or dropped packets.

The sequence number `mSequence field` is incremented by one each time a packet is sent. This approach allows the receiver to deduce lost and duplicate packets by identifying missing and duplicate sequence numbers.

The mSendTime field contains the time at which the packet was sent, represented as microseconds since January 1, 1970 UTC. This field is compared to the time the packet arrived at the receiver to determine the delay between the sender and the receiver.

In some embodiments of the invention, a plurality of one or more packets are sent over a path continuously. In some embodiments of the invention, the continuous stream of packet is denoted as a measurement stream. Each measurement stream is uniquely identified by the source and destination IP addresses. The sender maintains one socket descriptor for each source IP address it sends from and writes the destination IP address into the mDstAddr field. On the receiver side, the source IP address is returned by the recv() system call and the destination address is retrieved from the measurement packet.

Data Included in the Measurement Packets

In measurement packets that contain sufficient space, data will be included, including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

Some embodiments of the invention will add a single type of data to each packet. Some embodiments of the invention will use a complex data, including subpackets.

Some embodiments of the invention use subpackets that include a single byte subpacket type identifier, followed by a 2-byte length field (including the length of the type and length fields) and finally including the data that is to be sent. One embodiment will store all values in network byte order. Other byte orders will be apparent to those skilled in the art. The following data structure definition describes some embodiments.

```
class SubPacket {
  /*
   * The type identifier for this subpacket.
   */
  uint8_t mType;

  /*
   * The length of this subpacket, in network byte order.
   */
  uint16_t mLength;
};
```

One embodiment of this invention will include data describing a momentary snapshot of the measurement statistics for a given path between a sender and a receiver.

In some embodiments of this invention, this data will include one or more of the following information: the source and destination IP addresses that define the path, a measurement packet size for which the statistics have been calculated as well as computed measurement statistics that are at least partly

responsive to delay; computed measurement statistics that are at least partly responsive to jitter and computed measurement statistics that are at least partly responsive to packet loss.

In one embodiment of this invention, these statistics will be in units of microseconds expressed as 64-bit floating-point quantities and transmitted in a standard network byte order.

In one embodiment of this invention, the following data structure will store the computed statistics:

```
class TunnelStatsSubPacket : public SubPacket {
  /**
   * The time that this statistic snapshot was taken (in
   * microseconds since 1970).
   **/
  uint64_t mTimestamp;

  /**
   * The source IP address of the tunnel these statistics
apply
   * to.
   **/
  uint32_t mSrcAddr;

  /**
   * The destination IP address of the tunnel these
statistics
   * apply to.
   **/
  uint32_t mDstAddr;

  /**
   * The size of measurement packet that these statistics
apply
   * to.  A size of 0 indicates that these statistics
apply to
   * all packet sizes.
   **/
  uint16_t mPktSize;
```

```
    /**
     * The average delay in microseconds.
     **/
    double mDelay;

    /**
     * The average jitter in microseconds.
     **/
    double mJitter;

    /**
     * The percentage of packets that have been lost, in the
range
     * 0 to 1.
     **/
    double mLoss;
};
```

Some embodiments of this invention include the time at which the statistics were computed such that those statistics are sent over multiple paths for improved reliability and to take advantage of one path having less delay than another. One embodiment at the receiving end is able to compare the computation times of received statistics to place them in their original temporal order, regardless of their relative arrival times over the paths.

Some embodiments of this invention will send computed statistics specific to the paths that are part of the plurality of one or more paths that are between the specific sender and receiver. Other embodiments will send additional computed statistics for paths that are not one of the plurality of one or more paths that are between the specific sender and receiver.

Some embodiments of this invention will include network information concerning network topology including but not limited to information retrieved from routers such as in-bound or out-bound link utilization, inbound or out-bound link bandwidth and/or CPU utilization. Other network information determined from routers and other network devices will be apparent to someone skilled in the art.

Some embodiments of this invention will also include control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

In some embodiments of the invention, the control data will instruct a receiver to alter its configuration, including but not limited to zero or more of the following examples: instructing a receiver to initiate sending a plurality of one or more measurement packets, change one or more of the measurement packet sizes, inter-measurement packet transmission times and mix of packet sizes, and stop sending one or more of the plurality of measurement packets.

In some embodiments of the invention, this control information will include notification of measurement devices that have joined or left the network.

In many embodiments of the invention, the measurement packets will be encrypted by the sender and decrypted by the receiver. Some of these embodiments will use IPSec.

In some embodiments of the invention, the encryption and decryption will be done by an external device using IPSec.

Other encryption and decryption options will be apparent to one skilled in the art.

In some embodiments of the invention, the measurement packets will be digitally signed.

In some embodiments of the invention, a generic communication channel will be used by a sender and a receiver to communicate data between them.

Performance Characteristics of a Path

Measurements are used to compute performance characteristics of the paths traversed by the measurement packets. The measurements can either be computed from the measurement packets themselves, or extracted from the arbitrary data carried by the measurement packets. The measurements of performance characteristics include at least one or more of one-way measurements and round-trip measurements. The performance characteristics include at least one or more reachability, delay, jitter, loss, available bandwidth,

and total bandwidth. Other performance characteristics will be apparent to those skilled in the art.

In some embodiments of the invention, delay measurements are computed as the interval of time from the moment the measurement packet is sent by the sender to the moment of time the measurement packet is received by the receiver. The sending time is carried by the packet, and it is measured by the clock the sender refers to. The receiving time is measured by a clock that the receiver refers to, which may or may not be synchronized with the sender's clock.

In some embodiments of the invention, the clock of the sender and the clock of the receiver are synchronized. A plurality of one or more precise clock inputs such as GPS, NTP, IRIG and NIST will be used. Some embodiments of this invention will use the same clock as an input to more than one of the plurality of one or more senders and receivers. In some embodiments of the invention, the clock of the sender and the clock of the receiver are the same.

In some embodiments of the invention, the clock of the sender and the clock of the receiver are not synchronized, and mechanisms based on the measurement data are used to correct the clock skew and clock drift, the mechanisms including using minimum delay across multiple measurement samples, and using a mechanism to track the minimum delay over time.

Some embodiments of the invention will use the minimum round-trip delay between the devices to place a lower bound on clock skew.

Some embodiments of the invention will use the lower bound of multiple paths between the sender and receiver to further reduce the lower bound.

Some embodiments of the invention will correct for clock drift by tracking the relative clock skew between the sender and receiver over time and adjusting for the slope of the drift.

In some embodiments of the invention, jitter measurements, also known as inter-measurement packet delay variations, are computed as the difference in delay on consecutive, successfully received packets.

In some embodiments of the invention, jitter can also be computed as the difference between the instantaneous delay of a packet, and the average delay of all the measurement packets previously received.

In some embodiments of the invention, loss measurements are computed by assigning a timeout value to each measurement packet that indicates the instant of time after which the measurement packet will be declared lost, if the packet has not arrived by that time. In some embodiments of the invention, the timeout value of a measurement packet can be computed with the transmission time of a previously received packet, an estimation of the inter-transmission time between measurement packet, and an estimation of the transmission delay of the measurement packet. In some embodiments of the invention, the inter-transmission time can be estimated if the receiver knows about the scheduling pattern of transmission of measurement packets. In some embodiments of the invention, the transmission delay of packet can be estimated based on delay and jitter performance characteristics.

Performance characteristics of a path could be the measurement themselves, or statistics on those measurements. In the statistics case, a dynamic algorithm is used to updates the statistics associated with a path with every new measurement obtained with the arrival of every new packet over the path.

In some embodiments of the invention, the algorithm computes statistics over the performance characteristics of the path.

In some embodiments of the invention, the statistics include averages, deviations, and variances. Other statistics will be apparent to those skilled in the art. In some embodiments of the invention, averages can be computed using a plurality of one or more techniques including a moving average, an average based on the Robbins-Moro estimator, a window-based average or a bucket-based average. Other techniques to compute averages will be apparent to those skilled in the art.

In some embodiments of the invention, the moving average is an exponentially moving average computed using a Robbins-Moro estimator. The Robbins-Moro stochastic approximation estimator finds a solution of the equation:

$$E[f(t) - x] = 0$$

where E is the expectation, f(t) a function and x the estimator. The general form of the solution is:

$$x(t) = x(t-1) + alpha * [f(t) - x(t - 1)] = (1 - alpha) * x(t-1) + alpha * f(t)$$

or, with $alpha = (1 - \mu)$,

$$x = \mu * x + (1 - \mu) * f$$

μ is the weight of the estimator, and determines the amount contributed to the average by the function.. In some embodiments of the invention, μ is constant. In some embodiments of the invention, μ is a dynamic value, whose value depends on the last value of the function f according to the formula:

$$\mu = e \wedge (-f / K)$$

where K is a constant that also determines the importance of the last value of f with respect to the current value of the estimator x.

In some embodiments of the invention, average delay can be computed using an exponentially moving average as follows,

$$d = \mu * d + (1 - \mu) * m$$

where d is the exponentially moving average of delay, m is the last delay sample, and μ is the weight of the moving average.

In some embodiments of the invention, average jitter can be computed using an exponentially moving average as follows,

$$v = \mu * v + (1 - \mu) * |d - m|$$

where v is the exponentially moving average of jitter, $|d - m|$ is the last sample of jitter, and μ is the weight of the average.

In some embodiments of the invention, average jitter can be computed using an exponentially moving average as follows,

$$v = \mu * v + (1 - \mu) * |m - m'|$$

Where v is the exponentially moving average of jitter, $|m - m'|$ is the last sample of jitter, m is the last delay sample, m' is the previous delay sample, and μ is the weight of the average.

In some embodiments of the invention, delay and jitter averages can be combined into a single value as follows:

$$l = d + M * v$$

Where d is the average delay, v is the average jitter and M is a constant.

In some embodiments of the invention, average loss can be computed using an exponentially moving average as follows,

$$p\text{-}hat = \mu * p\text{-}hat + (1 - \mu) * p$$

where p-hat is the moving average of the loss, p = {0 if packet is received, 1 is the packet is declared lost}, and μ is the weight of the exponentially moving average.

In some embodiments of the invention, μ is determined based on the notion of forgiveness against a single packet loss. The forgiveness period is the interval of time between the time the packet loss occurs and the time the average loss is forgiven. The forgiveness period can be either defined in units of time, or in number of packets if the rate of the monitoring flow is known. That is, the forgiveness period will end after n consecutive packets have been received after the loss, when these packets have been transmitted at a certain rate.

The value of the exponentially moving average after receiving the n packets is needed before μ can be determined, and this value is known as the

forgiveness threshold. In some embodiments of the invention, the forgiveness threshold is chosen arbitrarily. In some embodiments of the invention, the forgiveness threshold takes the value:

$$½ (1 - \mu)$$

This value is half of the value of the estimator after the singe loss occurs, and thus we call it the *half-life threshold.* Similarly, we also call the forgiveness period under this threshold the *half-life period.* The advantage of using a forgiveness threshold greater than zero is that issues related to host-dependent floating-point representations reaching that value are avoided.

In some embodiments of the invention, μ is computed by comparing the value of the estimator after *n* consecutive packet arrivals since the loss with the *half-life threshold*:

$$p\text{-}hat = (1- \mu) * \mu \wedge n < ½ (1 - \mu)$$

Given that n is known because is determined by the value of the *half-life period* and the transmission rate, μ is computed as:

$$\mu = exp ((ln ½) / n)$$

In some embodiments of the invention, two thresholds are defined, an upper threshold and a lower threshold. When the value of p-hat exceeds the upper threshold, the loss is not forgiven until enough measurement packets are received consecutively so that the value of p-hat gets below the lower threshold.

Other mechanisms to compute μ will be apparent to for those skilled in the art.

Path Description

In some embodiments of the invention, the path traversed by the measurement packets from the sender to the receiver is such that the path is at least partly implemented with at least one of a GRE tunnel, an IPSEC tunnel and IPonIP tunnel. Other path implementations using tunnel will be apparent for those skilled in the art.

In some embodiments of the invention, the path traversed by the measurement packets from the sender to the receiver is implemented with a virtual circuit, including a frame relay PVC, an ATM PVC or MPLS. Other path implementations using virtual circuits will be apparent for those skilled in the art.

Other path implementations will be apparent to those skilled in the art.

Internetwork Description

In some embodiments of the invention, the internetwork is implemented by a plurality of one or more subnetworks, including a plurality of one or more VPNs, a plurality of one or more BGP autonomous systems, a plurality of one or more local area networks, a plurality of one or metropolitan area networks, and a plurality of one or morewide area networks.

In some embodiments of the invention, the internetwork is implemented by an overlay network.

Other internetwork implementations will be apparent to those skilled in the art.

Packet Sizes and Transmission Times

In some embodiments of the invention, the measurement packets are of varying sizes, including 64, 256, 512, 1024, 1500 bytes.

In some embodiments of the invention, the size of the measurement packets is specified with an external API.

In some embodiments of the invention, the measurement packets are of a fixed size.

In some embodiments of the invention, the measurement packet sizes and times between measurement packets simulate the traffic pattern of a plurality of one or more applications

In some embodiments of the invention, traffic patterns correspond to voice applications, where the packets re of small size, e.g., 30 bytes, and the inter-transmission time between consecutive packets is constant, e.g., 10 ms. These examples do not limit the possible size values and inter-transmission time values.

In some embodiments of the invention, traffic patterns correspond to video applications, where the packets size is the largest permitted to be transmitted by an internetwork without being fragmented, and the inter-transmission time between consecutive packets varies depending on the spatial and temporal complexity of the video content being transmitted, the compression scheme, the encoding control scheme.

In some embodiments of the invention, traffic patterns correspond to the plurality of applications observed in an internetwork, including at least one or more of HTTP transactions, FTP downloads, IRC communications, NNTP exchanges, streaming video sessions, VoIP sessions, videoconferencing sessions and e-commerce transactions. Other types of applications will be apparent to those skilled in the art.

In some embodiments of the invention, the inter-measurement packet transmission times are of varying length.

In some embodiments of the invention, the inter-measurement packet transmission times are of fixed length.

In some embodiments of the invention, the inter-measurement packet transmission times specified with an external API.

In some embodiments of the invention, the length of the inter-measurement packet transmission times is randomized according to a distribution. In some embodiments of the invention, this distribution is based at least in part on a uniform distribution. In some embodiments of the invention, this distribution is based at least in part on an exponential distribution. In some embodiments of the invention, this distribution is based at least in part on a geometric distribution. Other distributions will be apparent to those skilled in the art.

In some embodiments of the invention, the length of the inter-measurement packet transmission times is provided by a table.

In some embodiments of the invention, the length of the inter-measurement packet transmission times is controlled by a scheduler. In some embodiments of the invention, the scheduler uses a priority queue, keyed on desired send time.

Other mechanisms to specify the inter-measurement packet transmission time will be apparent to those skilled in the art.

Other packet sizes and transmission times will be apparent to those skilled in the art.

Path Selection

It is possible that multiple alternative paths between a sender and a receiver are available through an internetwork at any given moment. Performance characteristics of each of these paths can be used to select a subset of the paths.

In some embodiments of the invention, the subset of the plurality of paths is selected based at least in part on at least one of: one or more of the measurement statistics from the measurement packet and one or more of the computed statistics.

In some embodiments of the invention, the selection of the subset of the plurality of paths is based at least partly on the position of paths in a ranking. In some embodiments of the invention, the ranking is at least partly based on one or more of the measurement statistics included as data in the measurement packet. In some embodiments of the invention the ranking is at least partly based on the computed statistics of the path. In some embodiments of the invention the ranking is implemented by using a comparison function to compare the paths, and by ordering the paths in a decreasing order. In some embodiments of the invention the ranking is implemented by using a comparison function to compare the paths, and by ordering the paths in an increasing order. Other ranking techniques will be apparent to those skilled in the art.

In some embodiments of the invention, the ranking is based on a single score associated to each path. In some embodiments of the invention, this score is denoted *Magic Score* (MS), and it is computed as follows:

$$MS = ML * MF$$

$$ML = d + M * v$$

$$MF = delta * p\text{-}hat + 1$$

where ML is the *Magic Latency*, a component of the MS obtained using delay and jitter respectively calculated with statistics; and MF is the *Magic scaling Factor* that multiplies the value of ML, and is computed based on loss statistics. M is a constant that takes several values, including 4, for example. MS can be seen as a scaled-up version of ML, and the scaling factor MF is a function of p-hat and delta, a constant. As p-hat not only reflects loss but also detects large delay spikes before they happen, p-hat can be seen as an indicator of the departure of the path from a "normal mode" operation, and thus the scaling factor is only applied when there are loss or spikes. The goal of MF is to differentiate between paths that have very similar delay characteristics, but with one having losses and the other not having them.

In some embodiments of the invention, ML is used as a delay indicator, given that jitter is accounted as an increase in delay. In contrast, MS, although a scaled version of ML, cannot be used to indicate delay, except when MF = 1 (p-hat = 0), which leads to MS = ML. That means the value of MS is useful not by itself but to compare it with the MSs of other tunnels.

In some embodiments of the invention, loss statistics can be used as a discriminator instead of a scaling factor. That is, p-hat can eliminate paths experimenting loss. Then, the remaining paths can be selected using MS = ML.

In some embodiments of the invention, the selection of a subset of paths is based on applying at least one or more thresholds to at least one of more of the statistics.

In some embodiments of the invention, a single threshold is used, and computed as a certain percentage of the highest score of the paths. In some embodiments of the invention, the threshold is determined by subtracting a fixed quantity to the highest score of the paths.

In some embodiments of the invention, the number of paths in the subset of paths is fixed. In some embodiments of the invention, this fixed number of paths N out of M paths is determined such that the probability of having loss in

(M - N) paths simultaneously is less than a certain threshold. In some embodiments of the invention, this probability is a binomial, with the assumption that all paths have the same probability of loss.

In some embodiments of the invention, the selection of the subset of the plurality of paths is based at least partly on a probability associated with each path. In some embodiments of the invention, the probability of each path is at least partly based on one or more of the measurement statistics included as data in the measurement packet.

In some embodiments of the invention, the probabilities of each path are equal.

In some embodiments of the invention, the selection of the subset of the plurality of paths is based at least partly on the cost of the path.

In some embodiments of the invention, the selection of the subset of the plurality of paths is based at least partly on the amount of bandwidth consumed over a period of time.

Other possibilities to compute path probabilities will be apparent to those skilled in the art.

Other mechanisms to select a subset of the paths will be apparent to those skilled in the art.

CLAIMS

What is claimed is:

1. A method for communicating data within measurement traffic, the method comprising:

sending a plurality of one or more measurement packets over a plurality of one or more paths, each of the plurality of one or more paths traversing at least a portion of an internetwork, and each of the plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet,

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

2. The method of claim 1, wherein the measurements of performance characteristics include one-way measurements.

3. The method of claim 1, wherein the data includes measurement statistics.

4. A method for communicating data within measurement traffic, the method comprising:

receiving a plurality of one or more measurement packets over a plurality of one or more paths, each of the plurality of one or more paths traversing at least a portion of an internetwork, and each of the plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet,

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

5. The method of claim 4, wherein the measurements of performance characteristics include one-way measurements.

6. The method of claim 4, wherein the data includes measurement statistics.

7. The method of claim 4, further comprising analyzing of the measurement packet based on a dynamic algorithm, the dynamic algorithm computing computed statistics on one or more of the measurements of performance characteristics of at least a portion of the path of the measurement packet.

8. The method of claim 7, wherein a subset of the plurality of one or more paths is selected based at least in part on at least one of: one or more of the measurement statistics from the measurement packet and one or more of the computed statistics.

9. A method for communicating data within measurement traffic, the method comprising:

sending a first plurality of one or more measurement packets over a first plurality of one or more paths, each of the first plurality of one or more paths traversing at least a portion of an internetwork, and each of the first plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet,

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing

a receiver of the measurement packet to change one or more configuration parameters of the receiver,

receiving a second plurality of one or more measurement packets over a second plurality of one or more paths, each of the second plurality of one or more paths traversing at least a portion of an internetwork, and each of the second plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet, and

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

10. The method of claim 9, wherein the measurements of performance characteristics include one-way measurements.

11. The method of claim 9, wherein the data includes measurement statistics.

12. The method of claim 11, wherein the measurement statistics are at least partly responsive to jitter.

13. The method of claim 11, wherein the measurement statistics are at least partly responsive to delay.

14. The method of claim 11, wherein the measurement statistics are at least partly responsive to loss.

15. The method of claim 9, wherein the data includes control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

16. The method of claim 9, wherein the data includes a generic communication channel.

17. The method of claim 9, wherein the data includes network information.

18. The method of claim 9, wherein the data is embedded in multiple measurement packets that are sent over multiple paths for improved communication performance, including redundancy and shorter transmission time.

19. The method of claim 9, wherein the measurement packets are at least one of encrypted and digitally signed.

20. The method of claim 9, wherein a clock referred to by a sender of the measurement packet and a clock referred to by the receiver of the measurement packet are synchronized, the synchronization methods including at least one or more of GPS, NTP, IRIG, and NIST.

21. The method of claim 9, further comprising analyzing of the measurement packet based on a dynamic algorithm, the dynamic algorithm computing computed statistics on one or more of the measurements of performance characteristics of at least a portion of the path of the measurement packet.

22. The method of claim 21, wherein the algorithm computes averages of the measurements, including at least one of a moving average, an average based on the Robbins-Moro estimator, a window-based average, and a bucket-based average.

23. The method of claim 21, wherein the algorithm is at least partly specified through an external API.

24. The method of claim 21, wherein the computed statistics are at least partly recomputed upon the arrival of every measurement packet.

25. The method of claim 21, wherein the computed statistics are at least partly on measurement statistics from the measurement packet.

26. The method of claim 21, wherein a subset of the plurality of one or more paths is selected based at least in part on at least one of: one or more of the measurement statistics from the measurement packet and one or more of the computed statistics.

27. The method of 26, wherein the selection of the subset of the plurality of one or more paths is based at least partly on the position of paths in a ranking.

28. The method of 27, wherein the ranking is at least partly based on one or more of the measurement statistics included as data in the measurement packet.

29. The method of 26, wherein the selection of the subset of the plurality of one or more paths is based at least partly on a probability associated with each path of the plurality of one or more paths.

30. The method of claim 29, wherein the probability of each path of the plurality of one or more paths is at least partly based on one or more of the measurement statistics included as data in the measurement packet.

31. The method of 26, wherein the selection of the subset of the plurality of one of more paths is based at least partly on applying one or more thresholds to at least one of the measurements statistics included as data in the measurement packet.

32. The method of claim 9, wherein measurement packets at least partly rely on UDP.

33. The method of claim 9, wherein at least one of the plurality of one or more paths is at least partly implemented with at least one of a GRE tunnel and an IPSEC tunnel.

34. The method of claim 9, wherein at least one of the plurality of one or more paths is at least partly implemented with at least one of a frame relay PVC, an ATM PVC, and MPLS.

35. The method of claim 9, wherein the internetwork is a plurality of one or more subnetworks, including at least one of a plurality of one or more VPNs; an overlay network; a plurality of one or more BGP autonomous systems; a plurality of one or more local area networks; a plurality of one or more metropolitan area networks; and a plurality of one or more or wide area networks.

36. The method of claim 9, wherein the measurement packet sizes and times between measurement packets simulate the traffic pattern of a plurality of one or more applications.

37. The method of claim 36, wherein the plurality of one of more applications includes voice applications.

38. The method of claim 36, wherein the plurality of one of more applications includes video applications.

39. A networking system, comprising:

a plurality of one or more devices communicating at least a first plurality of one or more measurement packets over a first plurality of one or more paths, each of the first plurality of one or more paths traversing at least a portion of an internetwork, and each of the first plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet, and

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

40. The networking system of claim 39, wherein the plurality of one or more devices includes a first sub-plurality of one or more devices, wherein the first sub-plurality of one or more devices sends one or more of the first plurality of one or more measurement packets.

41. The networking system of claim 40, wherein the plurality of one or more devices includes a second sub-plurality of one or more devices, wherein the second sub-plurality of one or more devices receives one or more of a second plurality of one or more measurement packets over a second plurality of one or more paths, each of the second plurality of one or more paths traversing at least a portion of the internetwork, each of the second plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet,

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

42. The networking system of claim 41, wherein at least one of the first plurality of one or more measurement packets and at least one of the second plurality of one or more measurement packets are the same packet.

43. The networking system of claim 39, wherein at least one of the plurality of one or more devices receives one or more of the first plurality of one or more measurement packets.

44. The networking system of claim 39, wherein the plurality of one or more devices includes a first sub-plurality of one or more devices, wherein the first sub-plurality of one or more devices receives one or more of a second plurality of one or more measurement packets over a second plurality of one or more paths and sends one or more of the first plurality of one or more measurement packets, each of the second plurality of one or more paths traversing at least a portion of the internetwork, each of the second plurality of one or more measurement packets including:

information for a receiver of the measurement packet to compute measurements of performance characteristics of at least a portion of the path of the measurement packet,

data including one or more of measurement statistics, a generic communication channel, network information, and control data directing a receiver of the measurement packet to change one or more configuration parameters of the receiver.

45. The networking system of claim 44, wherein at least one of the first plurality of one or more measurement packets and at least one of the second plurality of one or more measurement packets are the same packet.

46. The networking system of claim 44, wherein the plurality of one or more devices includes a second sub-plurality of one or more devices, wherein the second sub-plurality of one or more devices sends one or more of the first plurality of one or more measurement packets.

47. The networking system of claim 46, wherein at least one of the first plurality of one or more measurement packets and at least one of the second plurality of one or more measurement packets are the same packet.

48. The networking system of claim 44, wherein the plurality of one or more devices includes a second sub-plurality of one or more devices, wherein the second sub-plurality of one or more devices receives one or more of the second plurality of one or more measurement packets.

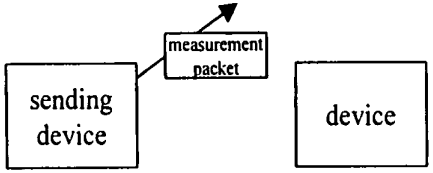
5

49. The networking system of claim 48, wherein at least one of the first plurality of one or more measurement packets and at least one of the second plurality of one or more measurement packets are the same packet.

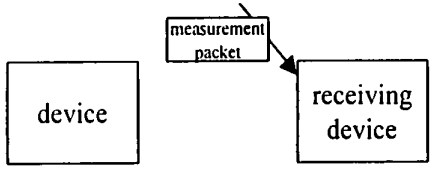
ABSTRACT

Methods and apparatuses for communicating data within measurement traffic are described. Embodiments that send, receive and both send and receive are described. Some embodiments are described that compute statistics based at least partly on measurement traffic. Some embodiments are described that communicate computed statistics within measurement traffic. Some embodiments are described that rank and select paths based at least in part on computed statistics.

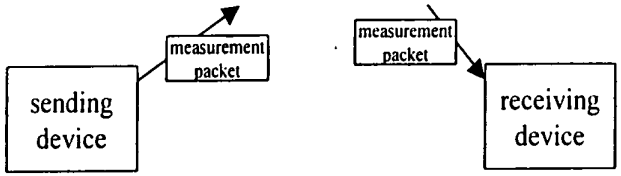
Example 1 – system with sending devices



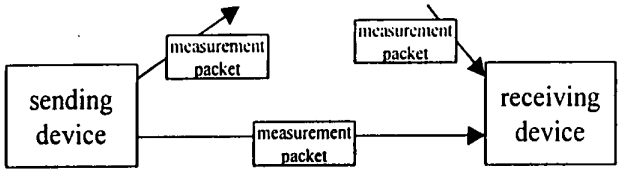
Example 2 – system with receiving devices



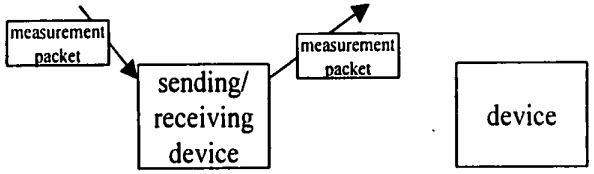
Example 3 – system with sending devices and receiving devices



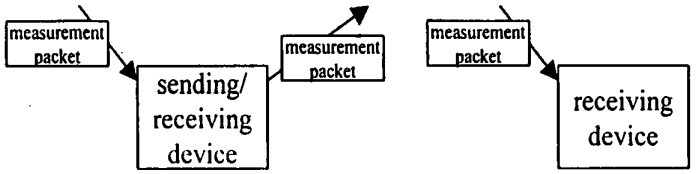
Example 4 – system with sending devices sending packets to receiving devices



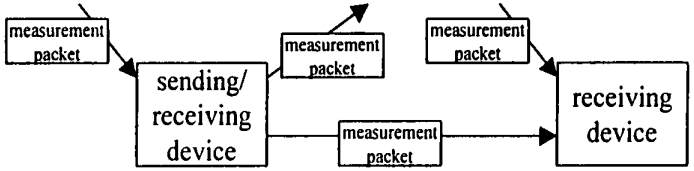
Example 5 – system with sending/receiving devices



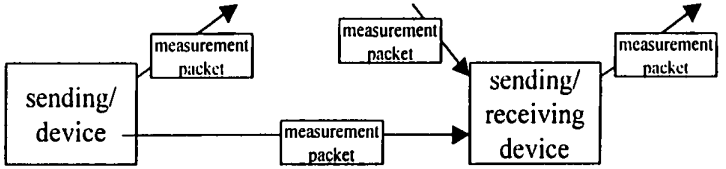
Example 6 – system with sending/receiving devices and receiving devices



Example 7 – system with sending/receiving devices sending packets to receiving devices



Example 8 – system with sending/receiving devices receiving packets from sending devices



Example 9 – system with sending/receiving devices sending packets to sending/receiving devices

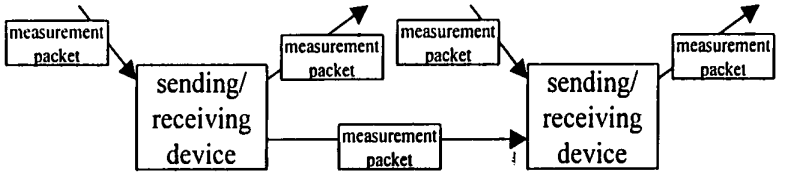


Figure 1

10/070339

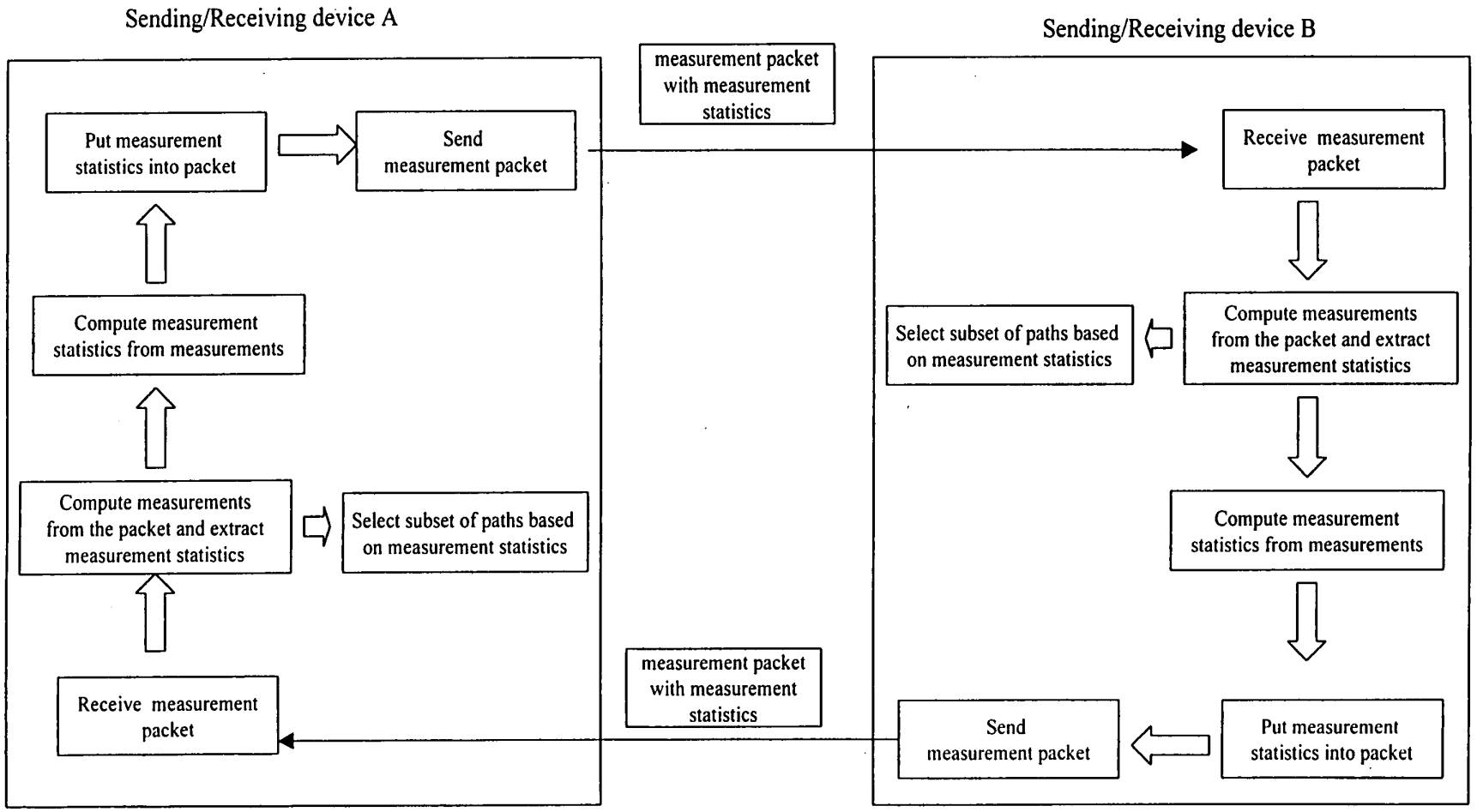


Figure 2

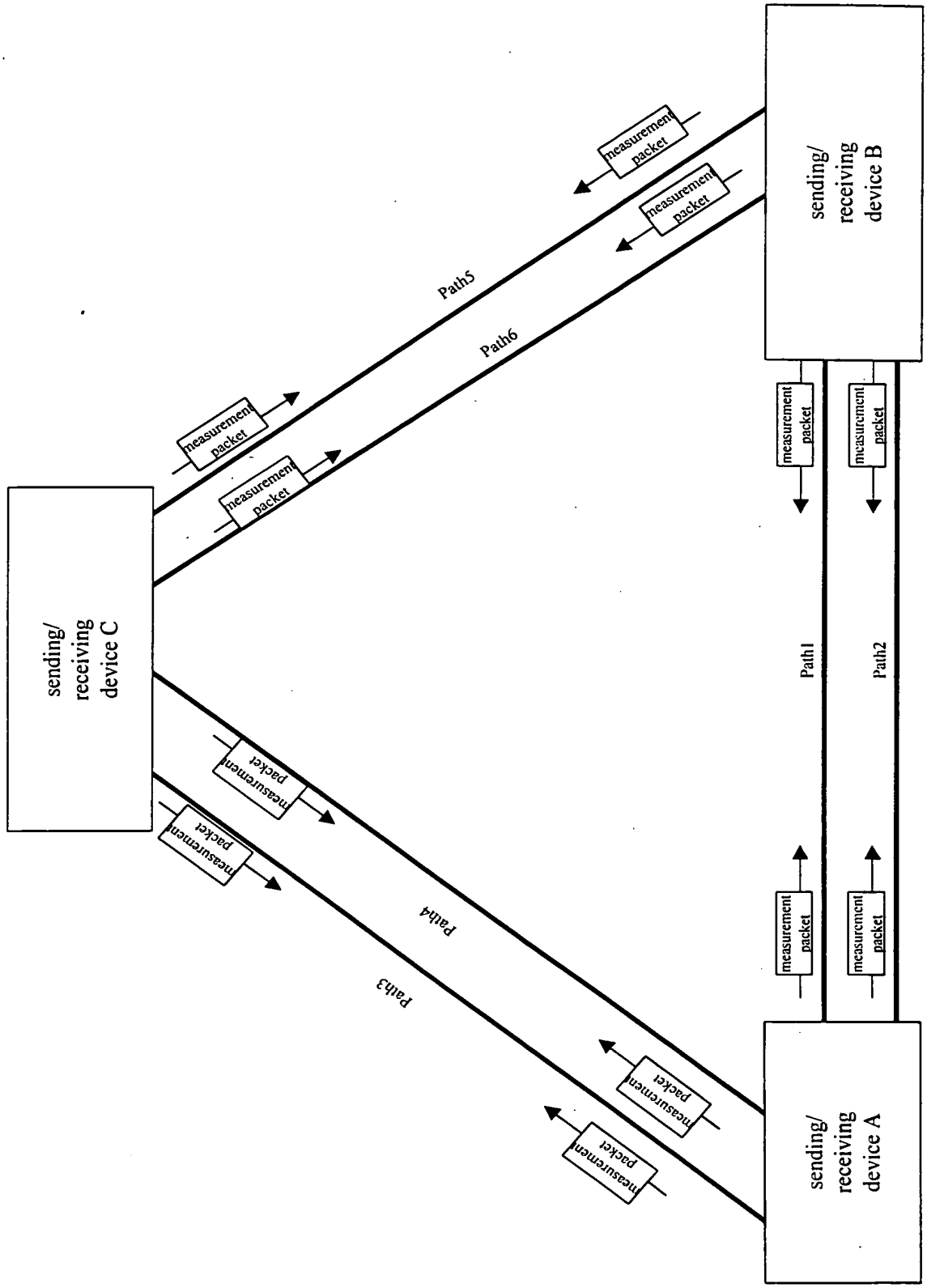

sending/
receiving
device A
sending/
receiving
device B
sending/
receiving
device C
Path1
Path2
Path3
Path4
Path5
Path6
measurement
packet
measurement
packet
measurement
packet
measurement
packet
measurement
packet
measurement
packet
measurement
packet
measurement
packet

Figure 3

INTERNATIONAL SEARCH REPORT

International Application No
PCT/01/32309

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, PAJ, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 793 976 A (CHEN THOMAS M ET AL) 11 August 1998 (1998-08-11) abstract figures 1,2 column 4, line 40 -column 5, line 25 column 7, line 5 -column 7, line 16 column 7, line 50 -column 9, line 14 claim 1 --- -/--	1,2,4,5, 9,10, 15-18, 20,36, 37,39

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 September 2002	16/09/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Pereira, M

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/01/32309

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5793976	A	11-08-1998	CA	2250278 A1	09-10-1997
			EP	0892958 A1	27-01-1999
			JP	2000507779 T	20-06-2000
			WO	9737310 A1	09-10-1997
EP 0942560	A	15-09-1999	US	6370163 B1	09-04-2002
			EP	0942560 A2	15-09-1999
			JP	11331222 A	30-11-1999
US 5563875	A	08-10-1996	NONE		
EP 0788267	A	06-08-1997	US	5822520 A	13-10-1998
			EP	0788267 A2	06-08-1997
			JP	9326796 A	16-12-1997
EP 0528075	A	24-02-1993	EP	0528075 A1	24-02-1993
			AU	2078892 A	25-02-1993
			CA	2076329 A1	20-02-1993
			JP	5207017 A	13-08-1993
			US	5343463 A	30-08-1994
US 5668800	A	16-09-1997	NONE		